Identification of a systemic lupus erythematosus risk locus spanning the genes *ATG16L2, FCHSD2,* and *P2RY2* in Koreans

Christopher J. Lessard[1, 24], Satria Sajuthi [2, 24,] Jian Zhao[3, 24], Kwangwoo Kim[4, 24], John A. Ice[1], He Li[1,5], Hannah Ainsworth[2], Astrid Rasmussen[1], Jennifer A. Kelly[1], So-Young Bang[4], Young Bin Joo[4], Jeongim Choi[4], Hye-Soon Lee[4], Young Mo Kang[6], Chang-Hee Suh[7], Won Tae Chung[8], Soo-Kon Lee[9], Jung-Yoon Choe[10], Seung Cheol Shim[11], Ji Hee Oh[12], Young Jin Kim[12], Bok-Ghee Han[12], Nan Shen[13,14], Hwee Siew Howe[15], Edward K. Wakeland[16], Quan-Zhen Li[16], Patrick M. Gaffney[1], Marta E. Alarcón-Riquelme [1,17], Lindsey A. Criswell[18], Chaim O. Jacob[19], Robert Kimberly[20], Timothy J. Vyse[21], John B. Harley[22, 23], Kathy L. Sivils[1, 4, 25], Sang-Cheol Bae[4, 25], Carl D. Langefeld[2, 25], and Betty P. Tsao[3, 25, 26]

[1] Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA
[2] Center for Public Health Genomics and Department of Biostatistical Sciences, Wake Forest University Health Sciences, Winston-Salem, NC
[3] Division of Rheumatology, Department of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
[4] Department of Rheumatology, Hanyang University Hospital for Rheumatic Diseases, Seoul 133-792, Republic of Korea
[5] Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73014, USA
[6] Kyungpook National University Hospital, Daegu 700-721, Republic of Korea
[7] Ajou University Hospital, Suwon 443-380, Republic of Korea
[8] Dong-A University Hospital, Busan 602-715, Republic of Korea
[9] Department of Internal Medicine, Yonsei University College of Medicine, Seoul 120-749, Republic of Korea
[10] Department of Internal Medicine, Catholic University of Daegu School of Medicine, Daegu 705-718, Republic of Korea
[11] Daejeon Rheumatoid & Degenerative Arthritis Center, Chungnam National University Hospital, Daejeon 305-764, Republic of Korea
[12] Korea National Institute of Health, Osong 361-709, Republic of Korea
[13] Shanghai Institute of Rheumatology, Renji Hospital,
[14] Shanghai JiaoTong University School of Medicine, Shanghai, China
[15] Tan Tock Seng Hospital, Singapore, Singapore
[16] University of Texas Southwestern Medical Center, Dallas
[17] Centro de Genoḿica e Investigaciones Oncoloḱgicas, Pfizer-Universidad de Granada-Junta de Andalućıa, Granada 18100, Spain;
[18] Rosalind Russell Medical Research Center for Arthritis, University of California San Francisco, San Francisco, CA
[19] Department of Medicine, University of Southern California, Los Angeles, CA;
[20] Department of Medicine, Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL, USA
[21] Divisions of Genetics and Molecular Medicine and Immunology, Infection and

Inflammatory Disease, King's College London, London, UK
[22] Division of Rheumatology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH
[23] US Department of Veterans Affairs Medical Center, Cincinnati, OH

[24] These authors contributed equally to this manuscript
[25] These authors co-directed this manuscript

[26] Corresponding Author:
Betty P. Tsao, PhD
900 Veteran Avenue, Warren Hall Rm 14-224
Division of Rheumatology, Department of Medicine
University of California
Los Angeles, CA 90095, USA
Tel: +1 (310) 825-8906
Email: btsao@mednet.ucla.edu

Abstract:

Systemic lupus erythematosus (SLE) is a chronic, heterogeneous autoimmune disorder characterized by inflammation, loss of tolerance to self-antigens, and dysregulated interferon responses. In this study, we performed a genome-wide association scan to identify loci associated in 1174 Korean SLE cases and 4248 population controls. Ten regions outside the HLA exceeded the genome-wide significance threshold of $P<5x10^{-8}$. Of these, 9 regions previously established SLE risk loci were replicated: *STAT1-STAT4, TNFSF4, TNFAIP3, IKZF1, HIP1, IRF5, BLK, WDFY4, and ETS1*. SNP-SLE association was also identified peaking at rs11235667 located between *FCHSD2* and *P2RY2* ($P = 1.0x10^{-8}$, odds ratio (OR) = 0.59, 95% confidence interval (CI)=0.50-0.71) on a haplotype spanning 33kb upstream to *ATG16L2*. Replication was tested for rs11235667 in an independent set of 1,412 SLE cases and 1,163 population controls of Korean and Chinese ancestries resulting in $P_{meta\text{-}rep}=0.001$ and $P_{meta\text{-}overall}=6.67x10^{-11}$ (OR=0.63, 95% CI=0.55-0.72). Within the HLA region, association peaked within the Class II at rs116727542 with multiple independent effects. Classical HLA allele imputation identified HLA-DRB1*1501 and HLA-DQB1*0602 as most strongly associated with SLE. Of the established SLE risk loci, independent second effects not previously reported were observed in *TNFAIP3* and *TNFSF4*, and differences in the association for a putative causal variant were identified in the *WDFY4* region. Further studies are need to determine if any other suggestive loci are true SLE risk effects and to identify the causal variant(s) in the region of *ATG16L2, FCHSD2,* and *P2RY2.*

Introduction:

Systemic lupus erythematosus (SLE; [MIM152700]) is a chronic, heterogeneous autoimmune disease characterized by the loss of tolerance to self-antigens, dysregulated type I interferon (IFN) responses, and inflammation, often resulting in systemic end-organ damage[1]. Immune dysfunction of SLE involves both B and T lymphocytes of the adaptive immune system, together with elements of the innate immune system, including dendritic cells and the complement system[1]. The clinical manifestations of SLE can be quite variable and can involve virtually any organ system. Although the precise etiology of SLE is largely unknown, the pathogenic mechanism likely involves environmental triggers in a genetically susceptible host[2]. Few effective treatment options exist, largely due to the lack of a complete understanding of the pathophysiological basis of the disease.

Genetic predisposition leading to an increased risk of SLE is supported by high heritability (>66%)[3], increased risk among siblings of affected patients ($\lambda_s \approx 30$)[4], and an ~25% monozygotic twin concordance[5; 6]. Today, associations of more than 50 loci with SLE susceptibility have been identified and confirmed[7]. Many of these genes fall into known pathways that are key to innate and adaptive immune responses, lymphocyte activation and/or function, and immune complex clearance[7]. However, a significant proportion of heritable risk to SLE has as of yet to be explained[8]. The identification of SLE-associated genes and their pathogenic mechanisms will greatly enhance our understanding of lupus pathophysiology and facilitate the development of effective diagnostic,

prognostic, and therapeutic tools. To date, large-scale genome-wide genetic studies of Asian SLE populations have focused on Han Chinese[9-11] and Japanese[12]. Moreover, several reports have shown that transracial mapping of SLE loci can aid in the dissection of risk effects[7]. In this study, we performed a genome-wide association (GWA) scan to identify genes associated with SLE in an East Asian population from Korea.

Methods:

Subjects:

A total of 1,174 patients with SLE were recruited from the Hanyang University Hospital for Rheumatic Diseases (HUHRD) and six other university hospitals in Korea[13]. In addition, 552 ethnically matched healthy controls were recruited from HUHRD. The 3,700 ethnically matched out-of-study population controls were recruited from the Korean National Institutes of Health[13]. In addition, an independent cohort of 1,412 SLE cases and 1,163 population controls were used for the replication studies[14; 15]. This sample set consisted of 739 Korean SLE cases and 436 Korean controls as well as 673 Chinese SLE cases and 727 Chinese controls (Supplementary Table 1).

Written, informed consent from each participant was obtained by each participant following protocols approved by the Institutional Review Boards of participating institutes. All cases used in this study fulfilled at least 4 of the 11 American College of Rheumatology criteria for SLE[16], while healthy, population-based controls were without family history of SLE or any other autoimmune

disease.

GWA scan Genotyping, Sample Quality Control, and Ascertainment of Populations Stratification:

Samples were genotyped using the Illumina HumanOmni1-Quad or HumanOmniExpress arrays using Infinium chemistry at Oklahoma Medical Research Foundation (OMRF) following the manufacturer's protocol (Illumina, Inc., San Diego, CA). The out-of-study GWA controls were genotyped on the HumanOmni1-Quad arrays by the Korea National Institutes of Health. Strict quality control standards were implemented for SNPs retained in the association analysis, including requirements for well-defined cluster scatter plots. Samples were excluded if they had a SNP call rate <90%. SNPs were considered high quality SNPs if they had call rates >95%, no evidence of differential missingness between cases and controls ($P < 0.05$) and no evidence of a departure from expected Hardy-Weinberg proportions (controls $P < 0.01$, cases $P < 0.000001$). Inference is primarily based on those SNPs with minor allele frequency (MAF) greater than 1%.

Based on the SNPs that passed the above quality control thresholds, samples were removed if there were inconsistencies between recorded and genotype-inferred gender or excess heterozygosity on the autosomes. Duplicates and first- or second-degree relatives were removed based on identity-by-descent statistics computed by the program KING[17]. Principal components were computed with the samples, merged with HapMap phase 3 individuals (CEU, YRI, and CHB) as reference populations[18] using EIGENSOFT[19]. PCA was

performed on a subset of autosomal SNPs that were selected by removing regions of known high linkage disequilibrium (LD), removing variants with MAF < 0.05, and pruning markers to reduce extended pairwise LD. Admixture estimates were computed using the program ADMIXTURE[20]. The principal components (PCs) and admixture estimates were used to remove genetic outliers (Supplementary Figure 1). The dataset that passed laboratory and statistical quality control was composed of 1174 SLE cases (1096 females and 78 males) and 548 within-study controls (547 females and 1 male). In addition, 3698 out-of-study controls (2330 females and 1368 males) were merged into the within-study genotype data.

Statistical Analysis:

To test for an association between a SNP and SLE status, a logistic regression analysis was computed including PC 1 as a covariate since no additional PC significantly changed the inflation factor (λ). Primary inference was based on the additive genetic model, unless there was significant lack-of-fit ($P < 0.05$). If there was evidence of a departure from an additive model, then inference was based on the most significant value from the dominant, additive or recessive genetic models. The additive and recessive models were computed only if there were at least 10 and 30 individuals homozygous for the minor allele, respectively. The analyses were completed using the program SNPGWA version 4.0. For the analysis of chromosome X SNPs, the samples were stratified by gender and then meta-analyzed across gender using the program METAL[21].

To determine the number of independent associations within each SLE-risk locus exceeding the genome-wide significance threshold, a manual stepwise model was computed. The stepwise modeling was implemented using the following procedure: forward selection with backward elimination using the entry and exit criteria of $P < 0.0001$ (accounting for approximately 500 independent variants within a given genomic region). Specifically, for each region of interest, the top SNP was included as a covariate and the association statistics were re-calculated. SNPs were allowed to enter and exit models in this stepwise fashion until no additional SNPs met a significance threshold of $P < 0.0001$.

Replication Genotyping, Sample Quality Control, and Ascertainment of Populations Stratification:

Genotypes were obtained using TaqMan assays (Life Technologies, Grand Island, NY) for four SNPs: rs2267828, rs10901656, rs11235667, and rs1048257. Analysis was conducted for these cohorts independently to allow for PC analysis using previously collected data. Ancestry adjustments for the Koreans were described previously in Lessard et al., 2012[14]. For the Chinese subjects, the PCA was done with slight modification from what was reported in Kaiser et al. 2013[15]. In this study, 7,918 randomly selected autosomal ImmunoChip SNPs with MAF>1%, low pairwise LD ($r^2 < 0.1$), and no evidence of association with SLE ($P > 0.01$) were used to perform PC analysis using EIGENSOFT. PC analysis plots of the CHB and CHS subjects in the 1000 Genomes Project along with our subjects were used to select and remove

genetic outliers. The first principal component (Chinese cohort) and PCs 1, 2, and 3 (OMRF and UCLA Korean datasets) were included as covariates in the logistic regression models based on the variance explained in each dataset. These dataset were than meta-analyzed using the program METAL[21]. To test for heterogeneity among the individual association results in the meta-analysis, we utilized both the Cochran's *Q* test statistic[22] and I^2 index[23].

Imputation:

To help localize the associations in the genome-wide significant regions, ungenotyped genetic markers were imputed based on the reference panel from the 1000 Genomes Project[24]. Specifically, the program SHAPEIT was used to pre-phase the genotype data[25]. After phasing the data, IMPUTE2 was used for the imputation with the 1000 Genomes Phase I integrated reference panel[26]. The imputed data was filtered using standard post-imputation quality control based on IMPUTE2 information scores >0.5 and confidence scores >0.9 for subsequent association tests. Post-association analysis required genotyped SNPs in LD with imputed variants to support the inferred alleles as true signals. The program SNPTESTv2 was used to test for association on the imputed data[27].

Imputation of the HLA classical alleles in the genes HLA A, B, C, DPB1, DQA1, DQB1 and DRB1 was done using the program HiBAG[28] and the Asian reference panel. In this sample, ~21% of the reference SNPs used by HiBAG were missing genotype data. To address this issue, HLA imputation was

repeated after filling-in the missing genotype data with the “best guess” imputed SNP data from the 1000 Genomes imputation described above. By using the “best guess” genotype data that had a posterior probability >0.90, the percent of missing SNPs in the reference set was reduced to 0.36%.

Results:

We observed a modest inflation in the test statistic (λ=1.09) with only slight deviation from expected once the HLA and other known SLE loci were removed (Supplementary Figure 2). A total of eleven regions surpassed the genome-wide significance threshold of $P < 5x10^{-8}$ with *STAT4* (MIM600558) yielding the most significant genotyped association with SLE at rs11889341 ($P = 8.02x10^{-19}$; Figure 1 and Table 1). Of the non-HLA regions, 9 risk loci had been previously identified and confirmed as risk loci for SLE, and included *STAT1* (MIM600555)-*STAT4*, *IKZF1* (MIM603023), *TNFAIP3* (MIM191163), *TNFSF4* (MIM603594), *HIP1* (MIM601767), *IRF5* (MIM607218), *ETS1* (MIM164740), *BLK* (MIM191305), and *WDFY4* (MIM613316). In addition, association not previously described for SLE risk was observed at 11q14 (Figure 1). This SNP-SLE association was observed with a single variant located between *FCHSD2* (MIM not available) and *P2RY2* (MIM600041) (rs11235667; $P=1.03x10^{-8}$; odds ratio (OR) = 0.59; 95% confidence interval (CI) = 0.50-0.71; Figure 2, Table 2, and Supplementary Table 2). Moreover, additional support was observed with genotyped variants in the region (Supplementary Table 2). After imputation of the 11q14 region showing association with SLE, rs11235667 remained the most significant association

(Figure 2 and Supplementary Table 2). However, a haplotype was identified with 8 variants exceeding the genome-wide significance threshold spanning from *ATG16L2* (MIM not available) through *FCHSD2* to the shared promoter region with *P2RY2*. Stepwise logistic regression analysis adjusting for rs11235667 indicated the presence of only a single effect (Supplementary Figure 3).

Replication analysis for the primary signal in the region of *FCHSD2-P2RY2* was done using an independent cohort from Korea and China (Supplementary Table 1). The SNP rs11235667 between *FCHSD2* and *P2RY2* continued to show significant SLE association and similar effect size ($P_{meta\text{-}rep}$ = 0.001; OR = 0.71, 95% CI = 0.57-0.87). The overall meta-analysis between GWA and replication studies yielded a $P_{meta\text{-}overall} = 6.67x10^{-11}$ (OR = 0.63, 95% CI = 0.55-0.72; Table 2). No evidence of heterogeneity was observed in the meta-analysis (Table 2).

Bioinformatics analysis using Haploreg v2[29] revealed that the region around rs11235667 was hypersensitive to DNase 1 in B cells by the ENCODE project[30]. This variant has been shown to be located within an enhancer element in multiple immunological cell types based on the Epigenetic Roadmap data (Supplementary Table 3)[31]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments carried out by the ENCODE project found POL2 and YY1 proteins cross-linked to this region. Moreover, sequence prediction methods indicate that rs11235667 can alter the binding motif for the FOXa family of transcription factors using sequence prediction methods according to Haploreg v2[29]. These data suggest the likely functional mechanism

involves regulation of expression of *ATG16L2, FCHSD2* and/or *P2RY2*. However, current eQTL databases do not suggest that rs11235667 influences the expression of these loci. This could be due to the lack of data in the correct cell and/or tissue type and/or that some databases do not interrogate this SNP in their studies.

Of the other 8 variants on the haplotype that exceeded genome-wide significance, rs11235604 had several findings that also make it intriguing as a potential causal variant (Supplemental Table 3). This variant is a missense allele (R220W) that resides in the coding region of *ATG16L2*. Although this polymorphism does change the amino acid sequence of the ATG16L2 protein, it is predicted to be benign by PolyPhen-2[32]. Haploreg v2[29] does report that rs11235604 alters 8 predicted regulatory motifs and is thought to be an active enhancer in several immunologically relevant cell types (Supplementary Table 3). Further work is needed to conclusively determine which polymorphism(s) are responsible for this association signal such as the following: evaluate the potential for rs11235667 impacts the expression of *ATG16L2, FCHSD2* and/or *P2RY2,* the impact of the missense mutation arising from rs11235604 on *ATG16L2*, and/or any other variant(s) within this haplotype.

One of the most consistent associations with SLE has been with the human leukocyte antigen (HLA) region. Although the HLA was not the most statistically significant genotyped region, the SNP rs116727542 ($P = 6.15 \times 10^{-24}$; Table 3), which is located in a broad peak of association that was observed from HLA-DR (MIM142860) through -DQ (MIM146880), showed the strongest SNP-

SLE association after imputation (Figure 3, Table 3, and Supplementary Table 4). The interval between HLA-DR and -DQ has previously been implicated Koreans[33]. In an attempt to identify the number of independent effects in this complex region, we used the stepwise approach described above and found ten independent effects (see Supplementary Figure 4A for results of the stepwise regression analysis). The first four variants identified in the stepwise regression analysis, rs116727542, rs9273371, rs114653103, and rs115253455, are all located in the HLA Class II region (Figure 3, Supplementary Figure 4A, and Supplementary Table 4).

To better understand the relationship between the variants reported in this study and the classical HLA alleles, we imputed alleles at HLA A, B, C, DPB1, DQA1, DQB1 and DRB1. The peak statistical significance was observed at $P = 5.55 \times 10^{-16}$ for both HLA-DRB1*1501 (OR = 1.85; 95% CI = 1.59-2.14) and HLA-DQB1*0602 (OR = 1.90; 95% CI = 1.62-2.21; Figure 3, Table 4, and Supplementary Table 5). Stepwise logistic regression modeling of the HiBAG-imputed HLA alleles identified ten independent effects (Table 4). To better relate the classical alleles to the variants identified in this GWA scan, stepwise modeling was done with both SNPs and classical HLA alleles. The peak effect after 1000 Genomes imputation of SNPs, rs116727542, was accounted for by HLA-DQB1*0602 and HLA-DRB1*0803 (see Supplementary Figure 4B for results of stepwise regression analysis).

Several previously identified non-HLA SLE loci were also replicated in this study, specifically the following: *STAT1-STAT4, TNFSF4, TNFAIP3, IKZF1,*

HIP1, IRF5, BLK, WDFY4, and ETS1 (Table 1, Figure 4, Supplementary Figures 5 to 10, and Supplementary Tables 6 to 14). Of these loci, the associations in the region of *TNFAIP3, TNFSF4*, and *WDFY4* have notable differences from previous studies.

In the region of *TNFAIP3* after imputation, the primary independent effect in the stepwise model was observed at rs5029937 (located within the second Intron of *TNFAIP3* (Table 1 and Figure 4A). The second independent effect was identified at rs9373203 around 3' of the *TNFAIP3* coding region. A previous SLE study in Han Chinese[9] reported rs2230926, and another SLE transracial mapping study in Koreans (with partial overlap of subjects with the current study) and Europeans[34] identified risk of SLE with rs7749323. Both variants (rs2230926 and rs7749323) are highly correlated with rs5029937, with D' = 1.0 and $r^2 > 0.98$, indicating our result is consistent with these previous reports (Supplementary Figure 11). In addition, the second effect here tagged by rs9373203 was not identified in either the Han et al 2009 or Adrianto et al. 2011 (Figure 4A). Musone et al. 2008[35] identified multiple effects with some spanning even further 3' of *TNFAIP3* than rs9373203. However, after their stepwise analysis, they identified rs6922466 as the tagging variant that accounted for this association; in the current study of Koreans; however, this variant is not associated with SLE. Moreover, the LD between these variants is very weak in Koreans ($r^2 = 0.00$; D' = 0.43), giving additional evidence that rs9373203 may be a true independent effect warranting further study (Supplementary Figure 11).

In the region of *TNFSF4* after imputation, two independent effects were

observed in the stepwise model. The first effect, peaking at rs76413021, is located in the first intron of the TNFSF4 coding region (Table 1 and Figure 4B), and it is in linkage disequilibrium with rs2205960 (D'=0.98, r^2=0.94) and rs1234315 (D' = 0.97, r^2 = 0.48), variants from the Han Chinese GWA scan[9] (Supplementary Figure 12). Moreover, this effect is consistent with the results reported in Europeans[36]. The second independent effect, which is distinct from the previous studies, peaks at rs4916342 (Figure 4B). Neither the Han et al. 2009[9] nor Cunninghame Graham et al. 2008[36] manuscripts describe association signals as far 5' of *TNFSF4* as our observation here with the second independent effect tagged by rs4916342. In a transracial mapping study of this region by Manku et al. 2013[37] that included subjects from East Asia, a second independent effect was identified that is tagged by rs1234314. In our current study of Koreans, we found that rs1234314 was accounted for in the stepwise model by rs76413021, the first effect. Although rs1234314 and rs4916342 are located in the same general genomic location, the LD structure further supports that they are not the same genetic effect; however, all the risk variants are located on a single risk haplotype (Supplementary Figure 12). This suggests that risk alleles for both rs76413021 and rs4916342 are needed to confer susceptibility to disease.

In the region of WDFY4, the current study did not replicate rs877819, which has been previously reported to result in a down regulation of *WDFY4* by modifying a YY1 binding site (Table 1)[38]. However, the results for *WDFY4* in Koreans are consistent with two previous studies. First, the most statistically

significant association within this region was rs7097397, a coding variant resulting in the amino acid substitution R1816Q ($P = 2.10x10^{-9}$), previously reported by Yang et al. 2010[11]. Second, we also demonstrate association for rs1913517, which was identified previously by Han et al. 2009[9] ($P = 2.54x10^{-5}$; Table 1 and Figure 4C). Our haplotype and stepwise regression analysis indicated that there were two independent effects in the region, with rs7097397 accounting for the association observed at rs1913517, and rs10857631 tagging the second independent effect (Figure 4C and Supplementary Figures 13 and 14).

In total, 15 genotyped variants surpassed the suggestive threshold of $P < 2x10^{-6}$ and were considered for further replication (Supplementary Table 15). Replication was attempted for three additional variants located within *GTF2IRD1, DOCK1*, and *AHNAK2,* all having multiple genotyped variants showing suggestive significance and/or having been previously implicated in other related phenotypes (Table 2). Only rs2267828 near *GTF2IRD1* yielded a $P_{meta\text{-}rep} < 0.05$, but this variant did not surpass genome-wide significance after meta-analysis with the GWA scan (Table 2). The variant in the region of *AHNAK2,* rs1048257, was trending towards significance, while rs10901656 near *DOCK1* showed association in one replication dataset with the opposite allele (Table 2). Outside of the 9 regions previously reported SLE loci described above, we observed only seven additional loci with $P < 5x10^{-5}$ but $P > 5x10^{-8}$ on the list of ~50 that have been described previously (Supplemental Table 16). This is likely due to the limited power of this study and/or a result of population-specific differences from

where the discoveries were originally identified.

Discussion:

The association in this region peaks between three candidate genes, *ATG16L2*, *FCHSD2*, and *P2RY2*, all of which have the biological potential to impact SLE pathophysiology. While this locus has not been reported in other systemic autoimmune diseases, variants in this region are associated with Crohn's disease (MIM266600) in Korean subjects[39]. Moreover, the peak variant identified in Crohn's disease, rs11235667, was also the variant discovered here in this current study with SLE. The missense variant, rs11235604, was also reported to be associated in Crohn's disease[39].

ATG16L2 (autophagy related 16-like 2) is a homologue of the gene *ATG16L1* (MIM610767) that has been implicated as a risk locus for Crohn's disease in patients of European descent[40]. Both loci are involved in autophagy; however, not much is known about the role *ATG16L2* plays in the process. Interestingly, this pathway has been implicated in SLE previously. The gene *ATG5* (MIM604261) has been previously implicated to be a risk locus for lupus[9; 41]. Studies in the mouse have shown that Apg16l (the mouse equivalent of human ATG16L) interacts with Apg5 (the mouse equivalent of human ATG5) suggesting that it is possible that ATG16L2 and ATG5 may interact too in the human[42]. More studies are needed to understand the function *ATG16L2* and if it is involved in this association signal.

FCHSD2 (FCH and double SH3 domains 2) has been described as

regulator of F-actin assembly through interactions with WAS (also known as WASP) and WASL (also known as N-WASP)[43]. *FCHSD2* is primarily expressed in CD19+ B cells and dendritic cells. Previous studies have shown that WAS plays an important role in the migration of T cells through reorganization of the actin cytoskeleton subsequent to interactions with dendritic or B cells[44].

P2RY2 (purinergic receptor P2Y, G-protein coupled, 2) is known to be involved in many cellular functions and to be expressed in monocytes and myeloid cells. P2RY2 is a receptor for ATP and UTP that acts as a sensor for the release of nucleotides by apoptotic cells[45]. Mice null for P2RY2 showed a decreased ability to recruit monocytes and macrophages upon activation of nucleotides from apoptotic cells[45]. P2RY2 is also known to induce CCL2 secretion in macrophages, and coding variants in the receptor have been shown to influence secretion of this proinflammatory chemokine[46].

Although the HLA region has been implicated in SLE susceptibility since the 1970s, the precise loci responsible for risk have not been fully characterized. A further cross comparison of populations will be beneficial to take advantage of differences in linkage disequilibrium likely will help further refine association signals seen in the GWA studies. For the classical alleles, previous studies have identified associations with alleles in the HLA-DR locus in Europeans, Chinese, Japanese, and Koreans, but HLA-DQB1*0602 has been implicated in Koreans before this current study[47-50]. Two prominent Classical HLA alleles identified in Europeans with SLE showed differences in association in Koreans. While HLA-DRB1*15:01 was among the most significantly associated with SLE, HLA-

DRB1*03:01 was found to be at very low frequency in this population and not associated with SLE (Supplementary Table 5). It is likely that there are several amino acid changing variants accounted for by the classical alleles that contribute to risk; however, more work is needed to determine if any non-coding RNAs or transcriptional changes also occur on these haplotypes.

This GWA scan replicated several loci that have been identified by prior studies, including *STAT1-STAT4, TNFSF4, TNFAIP3, IKZF1, HIP1, IRF5, BLK, WDFY4, and ETS1*. It is important to note that a previous GWA scan of Korean women with SLE has also reported replicating *STAT4* and *BLK* at a genome-wide significant level[33]. Although most of the signals in these loci are identical, we did describe a couple of notable differences with independent effects in *TNFSF4* and *TNFAIP3*. Moreover, we did not observe association with rs877819, which had been proposed as putative causal variant leading to expression differences[38].

In conclusion, we performed a GWA scan of Korean SLE cases and population controls in which we identified 11 regions that surpassed genome-wide significance. The ATG16L2 through *FCHSD2* to the promoter region of the *P2RY2* locus was identified and confirmed as an SLE-associated region. As has been widely reported since the 1970s, the HLA region was observed as a major risk factor for SLE. Here, we also show the relationship between the classical HLA alleles and the variants reported within this GWA study in Koreans. The nine additional regions, *STAT4, IKZF1, TNFAIP3, TNFSF4, HIP1, IRF5, ETS1, BLK,* and *WDFY4,* had previously been implicated in SLE. Additional replication

is needed for the suggestive loci identified in this study to determine their relationship with SLE. Although GWA approaches have been very successful in the identification of risk loci, continued efforts are need to narrow association signals to the causal variant(s) and to determine the functional causal mechanism(s) contributing to SLE pathogenesis.

Acknowledgments:

We are grateful to all the individuals with SLE and those serving as healthy controls who participated in this study. Sample collection and phenotyping of the subjects utilized in this study was made possible by funds from the Korea Healthcare Technology R&D Project of the Ministry for Health & Welfare (HI13C2124, S.C.B.). This publication was made possible by grants 5P01 AI083194 (C.J.L., A.R., P.M.G., M.E.A.R., L.A.C., C.O.J., R.P.K., T.J.V., J.B.H., C.D.L., K.L.S, and B.P.T.), 5P01 AR049084 (R.P.K. and C.D.L.), R01 AR043814 (B.P.T.), R21 AR065626 (B.P.T.) by the NIH. The contents are the sole responsibility of the authors and do not necessarily represent the official views of the NIH. Additional support was obtained from Wake Forest School of Medicine Center for Public Health Genomics (C.D.L.), Oklahoma Medical Research Foundation (C.J.L., P.M.G., and K.L.S.). The out-of-study Korean control data was provided from Korean Biobank Project supported by the Korea Center for Disease Control and Prevention at the Korea National Institute of Health.

Competing Financial Interests:

The authors declare no competing financial interests.

Supplemental Data Summary

Supplemental data includes 16 tables and 15 figures.

Web-based Resources:

OMIM, www.omim.org/

SNPGWA version 4.0, www.phs.wfubmc.edu

SHAPEIT, www.shapeit.fr/

References:

1. Tsokos, G.C. (2011). Systemic lupus erythematosus. N Engl J Med 365, 2110-2121.
2. Arbuckle, M.R., McClain, M.T., Rubertone, M.V., Scofield, R.H., Dennis, G.J., James, J.A., and Harley, J.B. (2003). Development of autoantibodies before the clinical onset of systemic lupus erythematosus. N Engl J Med 349, 1526-1533.
3. Lawrence, J.S., Martins, C.L., and Drake, G.L. (1987). A family survey of lupus erythematosus. 1. Heritability. J Rheumatol 14, 913-921.
4. Alarcon-Segovia, D., Alarcon-Riquelme, M.E., Cardiel, M.H., Caeiro, F., Massardo, L., Villa, A.R., and Pons-Estel, B.A. (2005). Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort. Arthritis Rheum 52, 1138-1147.
5. Block, S.R. (2006). A brief history of twins. Lupus 15, 61-64.
6. Deapen, D., Escalante, A., Weinrib, L., Horwitz, D., Bachman, B., Roy-Burman, P., Walker, A., and Mack, T.M. (1992). A revised estimate of twin concordance in systemic lupus erythematosus. Arthritis Rheum 35, 311-318.
7. Deng, Y., and Tsao, B.P. (2014). Advances in lupus genetics and epigenetics. Curr Opin Rheumatol 26, 482-492.
8. Dai, C., Deng, Y., Quinlan, A., Gaskin, F., Tsao, B.P., and Fu, S.M. (2014). Genetics of systemic lupus erythematosus: immune responses and end organ resistance to damage. Curr Opin Immunol 31C, 87-96.
9. Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H., Cai, Z.M., Huang, W., Zhao, G.P., et al. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet 41, 1234-1237.
10. Yang, W., Tang, H., Zhang, Y., Tang, X., Zhang, J., Sun, L., Yang, J., Cui, Y., Zhang, L., Hirankarn, N., et al. (2013). Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. Am J Hum Genet 92, 41-51.
11. Yang, W., Shen, N., Ye, D.Q., Liu, Q., Zhang, Y., Qian, X.X., Hirankarn, N., Ying, D., Pan, H.F., Mok, C.C., et al. (2010). Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. PLoS Genet 6, e1000841.
12. Okada, Y., Shimane, K., Kochi, Y., Tahira, T., Suzuki, A., Higasa, K., Takahashi, A., Horita, T., Atsumi, T., Ishii, T., et al. (2012). A genome-wide association study identified AFF1 as a susceptibility locus for systemic lupus eyrthematosus in Japanese. PLoS Genet 8, e1002455.
13. Kim, K., Bang, S.Y., Lee, H.S., Cho, S.K., Choi, C.B., Sung, Y.K., Kim, T.H., Jun, J.B., Yoo, D.H., Kang, Y.M., et al. (2014). High-density

genotyping of immune loci in Koreans and Europeans identifies eight new rheumatoid arthritis risk loci. Ann Rheum Dis 2015 Mar;74(3).
14. Lessard, C.J., Adrianto, I., Ice, J.A., Wiley, G.B., Kelly, J.A., Glenn, S.B., Adler, A.J., Li, H., Rasmussen, A., Williams, A.H., et al. (2012). Identification of IRF8, TMEM39A, and IKZF3-ZPBP2 as susceptibility loci for systemic lupus erythematosus in a large-scale multiracial replication study. Am J Hum Genet 90, 648-660.
15. Kaiser, R., Taylor, K.E., Deng, Y., Zhao, J., Li, Y., Nititham, J., Chang, M., Catanese, J., Begovich, A.B., Brown, E.E., et al. (2013). Brief Report: Single-nucleotide polymorphisms in VKORC1 are risk factors for systemic lupus erythematosus in Asians. Arthritis Rheum 65, 211-215.
16. Hochberg, M.C. (1997). Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 40, 1725.
17. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867-2873.
18. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.
19. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904-909.
20. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19, 1655-1664.
21. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-2191.
22. Cochran, W.G. (1954). The Combination of Estimates from Different Experiments. Biometrics 10, 101-129.
23. Higgins, J.P., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. Bmj 327, 557-560.
24. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
25. Delaneau, O., Marchini, J., and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. Nat Methods 9, 179-181.
26. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5, e1000529.
27. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet 11, 499-511.

28. Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R., and Weir, B.S. (2014). HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics J 14, 192-200.
29. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40, D930-934.
30. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
31. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.
32. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods 7, 248-249.
33. Lee, H.S., Kim, T., Bang, S.Y., Na, Y.J., Kim, I., Kim, K., Kim, J.H., Chung, Y.J., Shin, H.D., Kang, Y.M., et al. (2014). Ethnic specificity of lupus-associated loci identified in a genome-wide association study in Korean women. Ann Rheum Dis 73, 1240-1245.
34. Adrianto, I., Wen, F., Templeton, A., Wiley, G., King, J.B., Lessard, C.J., Bates, J.S., Hu, Y., Kelly, J.A., Kaufman, K.M., et al. (2011). Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat Genet 43, 253-258.
35. Musone, S.L., Taylor, K.E., Lu, T.T., Nititham, J., Ferreira, R.C., Ortmann, W., Shifrin, N., Petri, M.A., Kamboh, M.I., Manzi, S., et al. (2008). Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nat Genet 40, 1062-1064.
36. Cunninghame Graham, D.S., Graham, R.R., Manku, H., Wong, A.K., Whittaker, J.C., Gaffney, P.M., Moser, K.L., Rioux, J.D., Altshuler, D., Behrens, T.W., et al. (2008). Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. Nat Genet 40, 83-89.
37. Manku, H., Langefeld, C.D., Guerra, S.G., Malik, T.H., Alarcon-Riquelme, M., Anaya, J.M., Bae, S.C., Boackle, S.A., Brown, E.E., Criswell, L.A., et al. (2013). Trans-ancestral studies fine map the SLE-susceptibility locus TNFSF4. PLoS Genet 9, e1003554.
38. Zhao, H., Yang, W., Qiu, R., Li, J., Xin, Q., Wang, X., Feng, Y., Shan, S., Liu, Y., Gong, Y., et al. (2012). An intronic variant associated with systemic lupus erythematosus changes the binding affinity of Yinyang1 to downregulate WDFY4. Genes Immun 13, 536-542.
39. Yang, S.K., Hong, M., Zhao, W., Jung, Y., Baek, J., Tayebi, N., Kim, K.M., Ye, B.D., Kim, K.J., Park, S.H., et al. (2014). Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci

and common attributes of genetic susceptibility across ethnic populations. Gut 63, 80-87.

40. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J., et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 39, 207-211.

41. Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K., et al. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nat Genet 40, 204-210.

42. Mizushima, N., Kuma, A., Kobayashi, Y., Yamamoto, A., Matsubae, M., Takao, T., Natsume, T., Ohsumi, Y., and Yoshimori, T. (2003). Mouse Apg16L, a novel WD-repeat protein, targets to the autophagic isolation membrane with the Apg12-Apg5 conjugate. J Cell Sci 116, 1679-1688.

43. Cao, H., Yin, X., Cao, Y., Jin, Y., Wang, S., Kong, Y., Chen, Y., Gao, J., Heller, S., and Xu, Z. (2013). FCHSD1 and FCHSD2 are expressed in hair cell stereocilia and cuticular plate and regulate actin polymerization in vitro. PLoS One 8, e56516.

44. Lafouresse, F., Cotta-de-Almeida, V., Malet-Engra, G., Galy, A., Valitutti, S., and Dupre, L. (2012). Wiskott-Aldrich syndrome protein controls antigen-presenting cell-driven CD4+ T-cell motility by regulating adhesion to intercellular adhesion molecule-1. Immunology 137, 183-196.

45. Elliott, M.R., Chekeni, F.B., Trampont, P.C., Lazarowski, E.R., Kadl, A., Walk, S.F., Park, D., Woodson, R.I., Ostankovich, M., Sharma, P., et al. (2009). Nucleotides released by apoptotic cells act as a find-me signal to promote phagocytic clearance. Nature 461, 282-286.

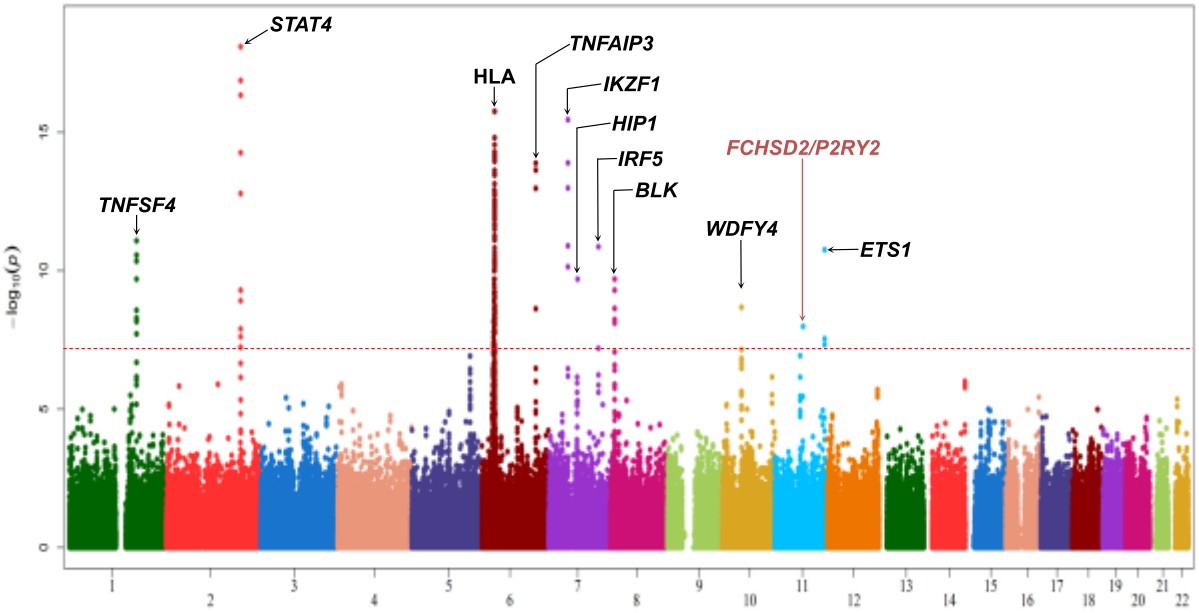
46. Higgins, K.R., Kovacevic, W., and Stokes, L. (2014). Nucleotides regulate secretion of the inflammatory chemokine CCL2 from human macrophages and monocytes. Mediators Inflamm 2014, 293925.

47. Lee, H.S., Chung, Y.H., Kim, T.G., Kim, T.H., Jun, J.B., Jung, S., Bae, S.C., and Yoo, D.H. (2003). Independent association of HLA-DR and FCgamma receptor polymorphisms in Korean patients with systemic lupus erythematosus. Rheumatology (Oxford) 42, 1501-1507.

48. Graham, R.R., Ortmann, W., Rodine, P., Espe, K., Langefeld, C., Lange, E., Williams, A., Beck, S., Kyogoku, C., Moser, K., et al. (2007). Specific combinations of HLA-DR2 and DR3 class II haplotypes contribute graded risk for disease susceptibility and autoantibodies in human SLE. Eur J Hum Genet 15, 823-830.

49. Furukawa, H., Oka, S., Shimada, K., Sugii, S., Hashimoto, A., Komiya, A., Fukui, N., Nagai, T., Hirohata, S., Setoguchi, K., et al. (2013). Association of increased frequencies of HLA-DPB1*05:01 with the presence of anti-Ro/SS-A and anti-La/SS-B antibodies in Japanese rheumatoid arthritis and systemic lupus erythematosus patients. PLoS One 8, e53910.

50. Furukawa, H., Kawasaki, A., Oka, S., Ito, I., Shimada, K., Sugii, S., Hashimoto, A., Komiya, A., Fukui, N., Kondo, Y., et al. (2014). Human leukocyte antigens and systemic lupus erythematosus: a protective role for the HLA-DR6 alleles DRB1*13:02 and *14:03. PLoS One 9, e87792.



1

2 **Figure 1. Summary of the genome-wide association results for 1174 SLE**

3 **cases and 3698 controls of Korean ancestry.** The -$\log_{10}$(P-value) for each

4 genotyped variant is plotted along the Y-axis with the chromosome and

5 chromosomal position along the X-axis. The red dash line indicates the genome-

6 wide significance threshold of P = 5×10^{-8}.

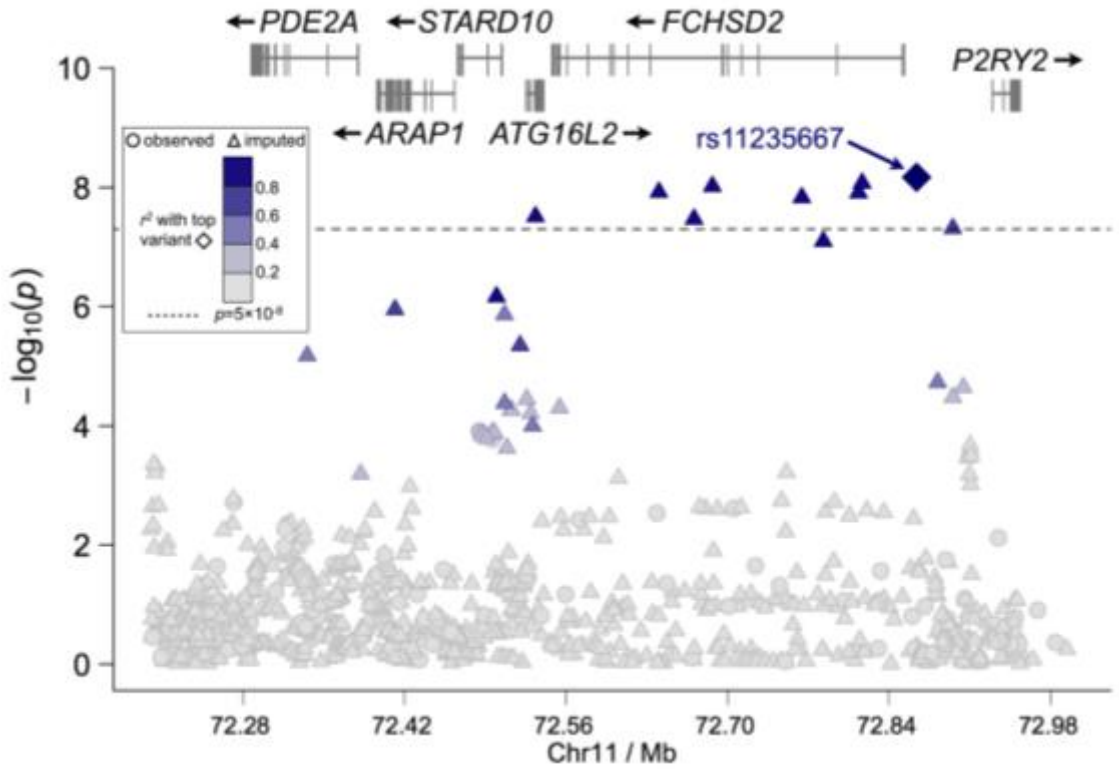


Figure 2. Zoomed plot of the region associated with SLE at 11q14. The -$\log_{10}$(P-value) is shown for each genotyped (shown as circles) and imputed (shown as triangles) variants are plotted with the peak association, rs11235667, plotted as a diamond. The linkage disequilibrium with rs11235667 is given by according to the scale on the figure. The genome-wide significance threshold is displayed as a dashed line at P = 5 x10^{-8}. Association exceeding this threshold was found extending from ATG16L2 through FCHSD2 to the shared promoter region with P2RY2.

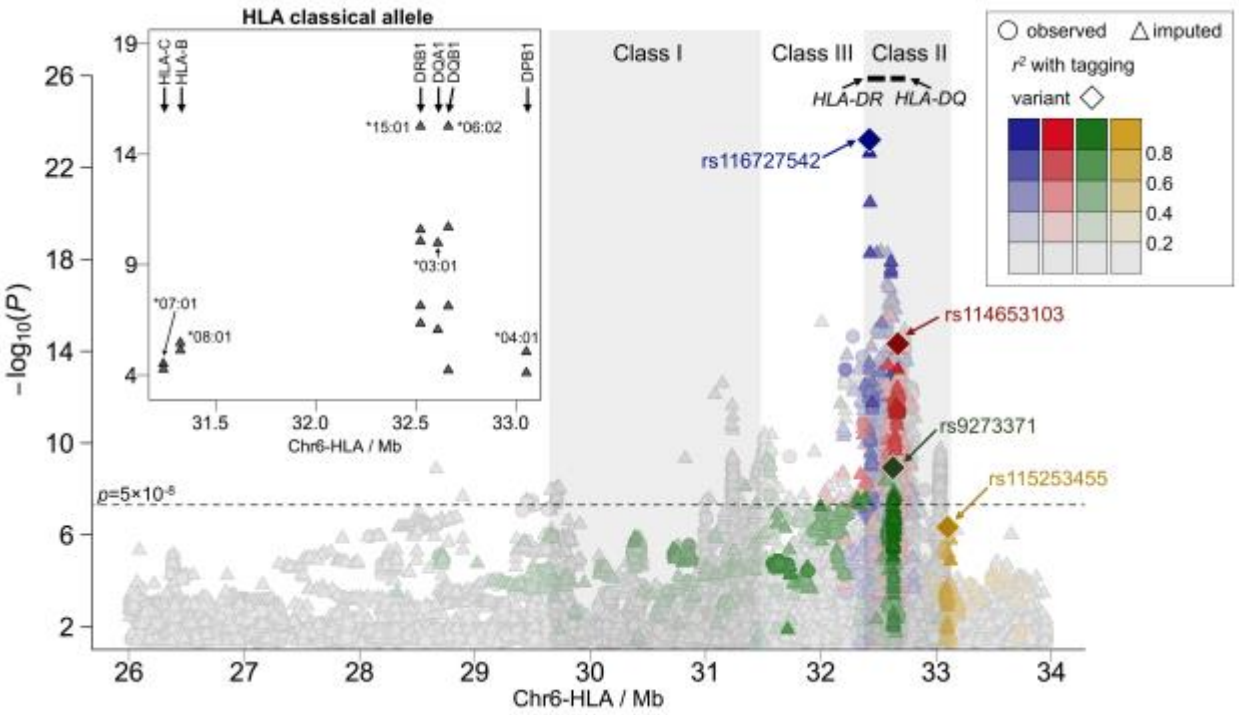


Figure 3. Expanded view of the association between SLE and the HLA region. The $-\log_{10}$(P-value) is shown for each observed (shown as circles) and imputed (shown as triangles) in the MHC region plotted in base pair position from 26Mb to 34Mb on chromosome 6. Linkage disequibrium with the first four variants included in the stepwise logistic regression analysis is shown with rs116727542 (blue diamond), rs114653103 (red diamond), rs9273371 (green diamond), and rs115253455 (gold diamond) all located within the HLA Class II region. The insert on the left shows the $-\log_{10}$(P-value) of the imputed classical alleles plotted according to base pair from 31Mb to 33.5Mb.

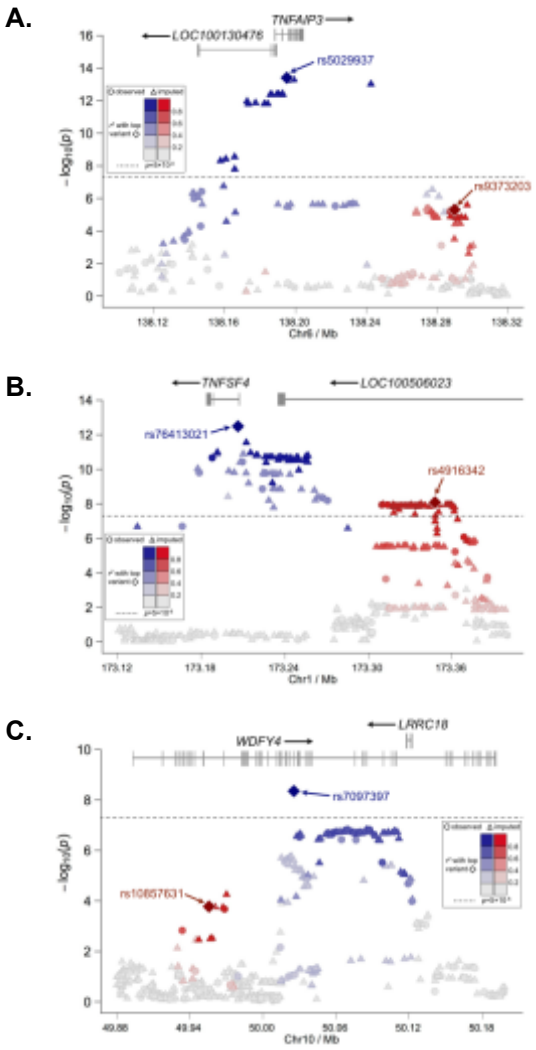


Figure 4. Expanded view of the associations with SLE with *TNFAIP3, TNFSF4,* and *WDFY4.* The -$\log_{10}$(P-value) is shown for each genotyped (shown as circles) and imputed (shown as triangles) variants are plotted for the region on Chromosome 6 in the region of *TNFAIP3* (A), Chromosome 1 for the *TNFSF4* (B) effects, and Chromosome 10 for *WDFY4* (C). For each independent effect, peak associations are represented by a diamond (blue for the first effect and red for the second, if applicable), and the correlation of variants accounted for by each effect is given in their respective color according the legends present in each plot. The genome-wide significance threshold is displayed as a dashed line on each plot at P = 5 x10^{-8}.

Table 1: Univaritae analysis of previous SLE associations outside of HLA region

Marker	Chr[A]	Position	Upstream Gene	Downstream Gene	Within Gene	MAF[B]	P_{GWAS}	Model	Obs/Imp[C]	OR (95% CI[D])	Maj/Min[E]
rs1234314	1	173177392	1kb from TNFSF4	269kb from PRDX6	-	0.37	$9.25x10^{-11}$	Add	Imp	1.37 (1.25-1.51)	C/G
rs2205960	1	173191475	15kb from TNFSF4	255kb from PRDX6	-	0.25	$1.03x10^{-11}$	Add	Imp	1.44 (1.30-1.60)	G/T
rs76413021	1	173206297	30kb from TNFSF4	240kb from PRDX6	-	0.23	$3.26x10^{-13}$	Add	Imp	1.52 (1.36-1.71)	G/A
rs844644	1	173209495	33kb from TNFSF4	237kb from PRDX6	-	0.4	$4.47x10^{-11}$	Add	Obs	1.37 (1.25-1.50)	A/C
rs10489265	1	173236065	60kb from TNFSF4	211kb from PRDX6	-	0.25	$8.19x10^{-12}$	Add	Obs	1.43 (1.29-1.58)	A/C
rs4916342	1	173347837	172kb from TNFSF4	99kb from PRDX6	-	0.3	$8.22x10^{-9}$	Add	Imp	0.75 (0.67-0.82)	A/G
rs16833239	2	191940260	-	-	STAT4	0.15	$9.69x10^{-10}$	Add	Imp	0.67 (0.59-0.76)	G/A
rs11889341	2	191943742	-	-	STAT4	0.32	$8.02x10^{-19}$	Add	Obs	1.53 (1.40-1.69)	C/T
rs12612769	2	191953998	-	-	STAT4	0.31	$2.37x10^{-19}$	Add	Imp	1.59 (1.43-1.75)	A/C
rs13192841	6	137967214	152kb from OLIG3	221kb from TNFAIP3	-	0.13	$7.50x10^{-3}$	Dom	Obs	1.23 (1.06-1.43)	G/A
rs5029937	6	138195151	-	-	TNFAIP3	0.07	$3.98x10^{-14}$	Dom	Imp	2.11 (1.74-2.55)	G/T
rs5029939	6	138195723	-	-	TNFAIP3	0.07	$4.09x10^{-14}$	Dom	Imp	2.10 (1.74-2.55)	C/G
rs2230926	6	138196066	-	-	TNFAIP3	0.07	$2.34x10^{-14}$	Dom	Obs	1.93 (1.63-2.28)	T/G
rs9373203	6	138289848	86kb from TNFAIP3	120kb from PERP	-	0.37	$4.89x10^{-6}$	Add	Imp	1.25 (1.14-1.37)	C/T
rs6922466	6	138444930	16kb from PERP	38kb from KIAA1244	-	0.18	0.297	Dom	Obs	1.08 (0.94-1.24)	A/G
rs11185602	7	50299077	100kb from C7orf72	45kb from IKZF1	-	0.33	$1.53x10^{-16}$	Add	Imp	0.66 (0.60-0.73)	A/G
rs17552904	7	50318308	120kb from C7orf72	26kb from IKZF1	-	0.33	$3.51x10^{-16}$	Add	Obs	0.65 (0.59-0.72)	G/T
rs6964720	7	75180344	-	-	HIP1	0.24	$2.00x10^{-10}$	Add	Obs	1.40 (1.26-1.56)	A/G
rs139110493	7	75209951	-	-	HIP1	0.06	$1.21x10^{-12}$	Dom	Imp	2.48 (1.93-3.19)	G/C
rs4728142	7	128573967	11kb from LOC392787	4kb from IRF5	-	0.14	$1.38x10^{-11}$	Add	Obs	1.53 (1.35-1.73)	G/A
rs113478424	7	128575797	13kb from LOC392787	2kb from IRF5	-	0.14	$3.97x10^{-12}$	Add	Imp	1.59 (1.39-1.81)	15-mer*/T
rs922483	8	11351912	-	-	BLK	0.25	$2.00x10^{-10}$	Add	Obs	0.71 (0.64-0.79)	T/C
rs2736345	8	11352485	-	-	BLK	0.24	$7.88x10^{-11}$	Add	Imp	0.70 (0.63-0.78)	G/A
rs10857631	10	49955821	-	-	WDFY4	0.13	$1.67x10^{-4}$	Add	Imp	1.30 (1.13-1.48)	A/G
rs7097397	10	50025396	-	-	WDFY4	0.37	$2.10x10^{-9}$	Add	Obs	1.33 (1.21-1.46)	A/G
rs877819	10	50042951	-	-	WDFY4	0.16	0.0558	Add	Imp	1.13 (1.00-1.28)	G/A
rs10776651	10	50084526	-	-	WDFY4	0.34	$1.54x10^{-7}$	Add	Imp	1.29 (1.18-1.43)	C/T
rs1913517	10	50119054	-	-	WDFY4	0.31	$2.54x10^{-5}$	Add	Obs	1.24 (1.12-1.37)	G/A
rs12576753	11	128304141	None within 500kb	25kb from ETS1	-	0.39	$1.74x10^{-11}$	Add	Obs	1.37 (1.25-1.56)	C/A
rs1128334	11	128328959	-	-	ETS1	0.38	$7.17x10^{-12}$	Add	Imp	1.39 (1.26-1.52)	G/A

[A] Chr = Chromosome

[B] MAF = Minor allele frequency

[C] Obs/Imp = observed/imputed

[D] CI = Confidence interval

[E] Maj/Min = Major/Minor allele

[*]15-mer= CTTAGCTATTGCTC

Table 2: Univariate analysis results for regions genotyped in the replication study.

Marker	Region Name	Maj/Min[A]	MAF[B] Case / Ctrl[C]	P_{GWAS}	Model	OR (95% CI[D])	P Meta Rep[E]	Q / I^2 Meta Rep[E]	OR (95% CI[D]) Meta Rep[E]	P Meta Overall[F]	Q / I^2 Meta Overall[F]	OR (95%CI[D]) Meta Overall
rs2267828	GTF2IRD1	A/G	0.40 / 0.45	$7.02x10^{-7}$	Add	0.79 (0.72-0.87)	0.02	0.56 / 0	0.87 (0.77-0.98)	$6.46x10^{-8}$	0.41 / 0	0.81 (0.76-0.88)
rs10901656	DOCK1	C/T	0.27 / 0.23	$6.91x10^{-7}$	Dom	1.39 (1.22-1.58)	0.095	0.6 / 0	1.14 (0.98-1.31)	$9.56x10^{-6}$	0.23 / 28.72	1.21 (1.12-1.32)
rs11235667	FCHSD2-P2RY2	A/G	0.07 / 0.11	$1.03x10^{-8}$	Add	0.59 (0.50-0.71)	0.0014	0.29 / 18.43	0.71 (0.57-0.87)	$6.67x10^{-11}$	0.14 / 44.37	0.63 (0.55-0.72)
rs1048257	AHNAK2	T/C	0.34 / 0.39	$1.67x10^{-6}$	Add	0.79 (0.72-0.87)	0.086	0.29 / 17.06	0.90 (0.80-1.01)	$8.66x10^{-7}$	0.12 / 47.82	0.82 (0.76-0.89)

[A] Maj/Min = Major/Minor allele

[B] MAF = Minor allele frequency

[C] Case/Ctrl = Case/Control

[D] CI = Confidence interval

[E] Meta Rep = Meta-analysis for the replication

Table 3: Univariate and stepwise results for top ten independent HLA associations

Marker[A]	Position	Upstream Gene	Downstream Gene	Within Gene	MAF[B]	P-value	OR (95%CI[C])	Maj/Min[D]	Stepwise P-value	Stepwise OR (95%CI[C])
rs116727542	32421227	8.4kb from *HLA-DRA*	64kb from *HLA-DRB5*	-	0.1700	$6.15x10^{-24}$	0.53 (0.47-0.60)	G/A	$1.96x10^{-18}$	1.74 (1.54-1.97)
rs9273371	32626565	14kb from *HLA-DQA1*	675bp from *HLA-DQB1*	-	0.1000	$1.18x10^{-9}$	1.61 (1.38-1.87)	C/T	$8.43x10^{-5}$	1.38 (1.18-1.63)
rs114653103	32668846	34kb from *HLA-DQB1*	40kb from *HLA-DQA2*	-	0.1200	$7.31x10^{-15}$	0.57 (0.49-0.66)	G/T	$1.81x10^{-13}$	0.50 (0.41-0.60)
rs115253455	33100021	43kb from *HLA-DPB1*	30kb from *COL11A2*	-	0.1300	$4.68x10^{-7}$	0.70 (0.61-0.80)	T/A	$4.51x10^{-7}$	0.66 (0.56-0.77)
chr6:31996524	31996524	-	-	*C4B*	0.1700	$4.95x10^{-8}$	0.71 (0.63-0.80)	C/A	$4.87x10^{-10}$	0.64 (0.55-0.73)
rs113833333	32594898	37kb from *HLA-DRB1*	10kb from *HLA-DQA1*	-	0.4100	$2.30x10^{-5}$	0.82 (0.74-0.90)	C/T	$3.09x10^{-8}$	0.74 (0.67-0.82)
rs116427960	31319226	79kb from *HLA-C*	2.4kb from *HLA-B*	-	0.0081	$8.96x10^{-7}$	4.57 (2.49-8.38)	C/T	$1.69x10^{-5}$	3.17 (1.87-5.35)
rs114904515	29362756	20kb from *OR12D3*	1.7kb from *OR12D2*	-	0.1000	$1.55x10^{-6}$	0.69 (0.60-0.80)	C/T	$5.70x10^{-6}$	0.67 (0.56-0.79)
rs118044183	30954150	-	-	*MUC21*	0.1200	$3.61x10^{-6}$	1.50 (1.27-1.79)	C/T	$1.39x10^{-4}$	1.39 (1.17-1.64)
rs2736191	31560910	~150bp from *NCR3*	22kb from *AIF1*	-	0.3800	$7.53x10^{-7}$	0.79 (0.71-0.86)	C/G	$1.01x10^{-4}$	0.81 (0.73-0.90)

[A] All variants within this table have been imputed.

[B] CI = Confidence interval

[C] MAF = Minor allele frequency

[D] Maj/Min = Major allele/Minor allele

Note: Tables are in the order they were identified in the stepwise model. The stepwise results presented in this table are for adjusting for all other variants in the table. For complete results in the HLA region, please refer to Supplementary Table 4.

Table 4: Univariate and stepwise results of HIBAG-imputed HLA dosages

	Dosage Frequency		Best Guess count					
HLA Allele	**Cases**	**Controls**	**Cases**	**Controls**	**OR (95% CI[A])**	**P-value**	**Stepwise P-value**	**Stepwise OR (95%CI[A])**
DQB1*06:02[B]	0.25	0.15	301	636	1.90 (1.62 - 2.21)	5.55×10^{-16}	1.93×10^{-23}	2.35 (1.99 – 2.78)
DRB1*08:03	0.20	0.13	251	613	1.59 (1.34 - 1.88)	7.37×10^{-8}	7.63×10^{-16}	2.14 (1.78 - 2.58)
DQB1*02:02	0.18	0.12	212	502	1.60 (1.35 - 1.90)	7.57×10^{-8}	2.19×10^{-18}	2.50 (2.04 - 3.07)
DQA1*03:02	0.19	0.16	286	847	1.41 (1.15 - 1.75)	1.27×10^{-3}	4.92×10^{-9}	1.98 (1.58 - 2.50)
B*08:01	0.02	0.004	20	14	5.43 (2.66 - 11.08)	3.42×10^{-6}	5.15×10^{-6}	5.71 (2.70 -12.07)
DQA1*04:01	0.04	0.03	54	133	1.73 (1.17 - 2.57)	5.94×10^{-3}	3.11×10^{-5}	2.36 (1.58 - 3.54)
C*07:02	0.22	0.17	261	711	1.37 (1.17 - 1.59)	5.45×10^{-5}	2.06×10^{-3}	1.30 (1.10 - 1.54)
DRB1*04:06	0.03	0.07	41	354	0.15 (0.09 - 0.26)	2.63×10^{-11}	1.11×10^{-4}	0.32 (0.18 - 0.57)
DPB1*05:01	0.79	0.72	930	3092	1.16 (1.05 - 1.28)	2.87×10^{-3}	1.06×10^{-4}	1.23 (1.11 - 1.36)
DRB1*16:02	0.03	0.02	38	87	1.78 (1.16 - 2.72)	7.73×10^{-3}	1.53×10^{-3}	2.05 (1.32 - 3.20)
DPB1*17:01	0.03	0.04	41	160	0.90 (0.62 - 1.30)	5.66×10^{-1}	1.07×10^{-3}	0.5 (0.33 - 0.76)
C*01:02	0.29	0.32	359	1429	0.88 (0.77 - 1.01)	6.51×10^{-2}	5.82×10^{-3}	0.82 (0.71 - 0.94)
DRB1*12:02	0.04	0.07	39	266	0.52 (0.37 - 0.74)	2.98×10^{-4}	7.07×10^{-3}	0.61 (0.42 - 0.87)

[A] CI = Confidence interval

[B] The allele HLA-DRB1*15:01 had the same P-value as HLA-DQB1*06:02, but the later was selected by the stepwise modeling procedure. For complete results, please refer to Supplementary Table 5.

Note: Tables are in the order they were identified in the stepwise model. The stepwise results presented in this table are for adjusting for all other variants in the table.